Another issue is raised by the standard error results. Often we use the data to help determine the model. Once a model is built or selected, inferences and predictions may be made. Usually inferences are based on the assumption that the selected model was fixed in advance and so only reflect uncertainty concerning the parameters of that model. Students took that approach here. Because the uncertainty concerning the model itself is not allowed for, these inferences tend to be overly optimistic leading to unrealistically small standard errors. Methods for realistic inference when the data are used to select the model have come under the heading of *Model Uncertainty* — see Chatfield (1995) for a review. The effects of model uncertainty often overshadow the parametric uncertainty and the standard errors need to be inflated to reflect this. Faraway (1992) developed a bootstrap approach to compute these standard errors while Draper (1995) is an example of a Bayesian approach. These methods are a step in the right direction in that they reflect the uncertainty in model selection. Nevertheless, they do not address the problem of model multiplicity since they proscribe a particular method of analysis that does not allow for differences between human analysts.

Sometimes the data speak with a clear and unanimous voice — the conclusions are incontestable. Other times, differing conclusions may be drawn depending on the model chosen. We should acknowledge the possibility of alternative conflicting models and seek them.

CHAPTER 11

Insurance Redlining — A Complete Example

In this chapter, we present a relatively complete data analysis. The example is interesting because it illustrates several of the ambiguities and difficulties encountered in statistical practice.

Insurance redlining refers to the practice of refusing to issue insurance to certain types of people or within some geographic area. The name comes from the act of drawing a red line around an area on a map. Now few would quibble with an insurance company refusing to sell auto insurance to a frequent drunk driver, but other forms of discrimination would be unacceptable.

In the late 1970s, the U.S. Commission on Civil Rights examined charges by several Chicago community organizations that insurance companies were redlining their neighborhoods. Because comprehensive information about individuals being refused homeowners insurance was not available, the number of FAIR plan policies written and renewed in Chicago by zip code for the months of December 1977 through May 1978 was recorded. The FAIR plan was offered by the city of Chicago as a default policy to homeowners who had been rejected by the voluntary market. Information on other variables that might affect insurance writing such as fire and theft rates were also collected at the zip code level. The variables are:

race racial composition in percentage of minority

fire fires per 100 housing units

theft theft per 1000 population

age percentage of housing units built before 1939

involact new FAIR plan policies and renewals per 100 housing units

income median family income in thousands of dollars

side North or South Side of Chicago

The data come from Andrews and Herzberg (1985) where more details of the variables and the background are provided.

11.1 Ecological Correlation

Notice that we do not know the race of those denied insurance. We only know the racial composition in the corresponding zip code. This is an important difficulty that needs to be considered before starting the analysis.

When data are collected at the group level, we may observe a correlation between two variables. The ecological fallacy is concluding that the same correlation holds at the individual level. For example, in countries with higher fat intakes in the diet,

higher rates of breast cancer have been observed. Does this imply that individuals with high fat intakes are at a higher risk of breast cancer? Not necessarily. Relationships seen in observational data are subject to confounding, but even if this is allowed for, bias is caused by aggregating data. We consider an example taken from U.S. demographic data:

```
> data(eco)
> plot(income ~ usborn, data=eco, xlab="Proportion US born",
  ylab="Mean Annual Income")
```

In the first panel of Figure 11.1, we see the relationship between 1998 per capita income dollars from all sources and the proportion of legal state residents born in the United States in 1990 for each of the 50 states plus the District of Columbia (D.C.). We can see a clear negative correlation.

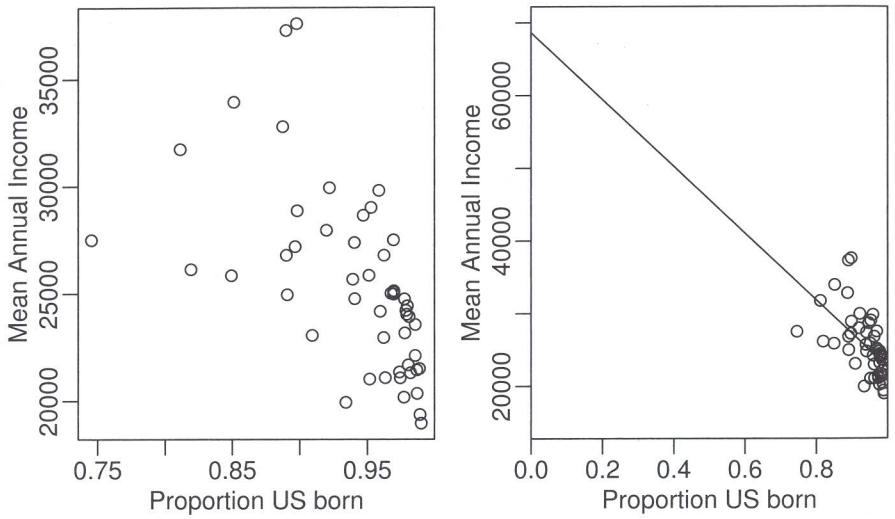

Figure 11.1 *1998 annual per capita income and proportion U.S. born for 50 states plus D.C. The plot on the right shows the same data as on the left, but with an extended scale and the least squares fit shown.*

We can fit a regression line and show the fitted line on an extended range:

```
> g <- lm(income ~ usborn, eco)
> summary(g)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    68642       8739    7.85  3.2e-10
usborn        -46019       9279   -4.96  8.9e-06

Residual standard error: 3490 on 49 degrees of freedom
Multiple R-Squared: 0.334,      Adjusted R-squared: 0.321
F-statistic:  24.6 on 1 and 49 DF,  p-value: 8.891e-06
```



```
> plot(income ~ usborn, data=eco, xlab="Proportion US born",
  ylab="Mean Annual Income",xlim=c(0,1),ylim=c(15000,70000),
  xaxs="i")
> abline(coef(g))
```

We see that there is a clear statistically significant relationship between the per capita annual income and the proportion who are U.S. born. What does this say about the average annual income of people who are U.S. born and those who are naturalized citizens? If we substitute `usborn=1` into the regression equation, we get 68642 – 46019 = $22,623, while if we put `usborn=0`, we get $68,642. This suggests that on average, naturalized citizens earn three times more than U.S. born citizens. In truth, information from the U.S. Bureau of the Census indicates that U.S. born citizens have an average income just slightly larger than naturalized citizens. What went wrong with our analysis?

The ecological inference from the aggregate data to the individuals requires an assumption of constancy. Explicitly, the assumption would be that the incomes of the native born do not depend on the proportion of native born within the state (and similarly for naturalized citizens). This assumption is unreasonable for these data because immigrants are naturally attracted to wealthier states.

This assumption is also relevant to the analysis of the Chicago Insurance data since we have only aggregate data. We must keep in mind that the results for the aggregated data may not hold true at the individual level.

11.2 Initial Data Analysis

Start by reading the data in and examining it:

```
> data(chredlin)
> chredlin
      race fire theft  age involact income side
60626 10.0  6.2    29 60.4      0.0 11.744    n
60640 22.2  9.5    44 76.5      0.1  9.323    n
...etc...
60645  3.1  4.9    27 46.6      0.0 13.731    n
```

Summarize:

```
> summary(chredlin)
      race            fire           theft            age
 Min.   : 1.00   Min.   : 2.00   Min.   :  3.0   Min.   : 2.0
 1st Qu.: 3.75   1st Qu.: 5.65   1st Qu.: 22.0   1st Qu.:48.6
 Median :24.50   Median :10.40   Median : 29.0   Median :65.0
 Mean   :34.99   Mean   :12.28   Mean   : 32.4   Mean   :60.3
 3rd Qu.:57.65   3rd Qu.:16.05   3rd Qu.: 38.0   3rd Qu.:77.3
 Max.   :99.70   Max.   :39.70   Max.   :147.0   Max.   :90.1
    involact         income       side
 Min.   :0.000   Min.   : 5.58   n:25
 1st Qu.:0.000   1st Qu.: 8.45   s:22
 Median :0.400   Median :10.69
```

```
 Mean   :0.615   Mean   :10.70
 3rd Qu.:0.900   3rd Qu.:11.99
 Max.   :2.200   Max.   :21.48
```

We see that there is a wide range in the `race` variable, with some zip codes almost entirely minority or nonminority. This is good for our analysis since it will reduce the variation in the regression coefficient for race, allowing us to assess this effect more accurately. If all the zip codes were homogeneous, we would never be able to discover an effect from these aggregated data. We also note some skewness in the `theft` and `income` variables. The response `involact` has a large number of zeros. This is not good for the assumptions of the linear model but we have little choice but to proceed. We will not use the information about North vs. South Side until later. Now make some graphical summaries:

```
> par(mfrow=c(2,3))
> for(i in 1:6) stripchart(chredlin[,i],main=names(chredlin)[i],
  vertical=TRUE,method="jitter")
> par(mfrow=c(1,1))
> pairs(chredlin)
```

The strip plots are seen in Figure 11.2. Jittering has been added to avoid overplotting of symbols. Now look at the relationship between `involact` and `race`:

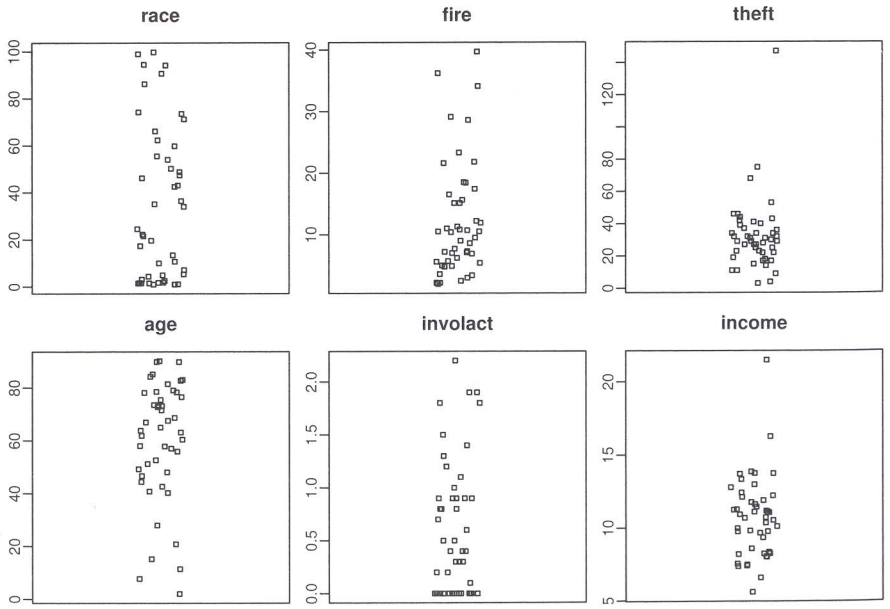

Figure 11.2 *Strip plots of the Chicago Insurance data.*

```
> summary(lm(involact ~ race,chredlin))
Coefficients:
```



```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12922    0.09661    1.34     0.19
race         0.01388    0.00203    6.84  1.8e-08

Residual standard error: 0.449 on 45 degrees of freedom
Multiple R-Squared: 0.509,      Adjusted R-squared: 0.499
F-statistic: 46.7 on 1 and 45 DF,  p-value: 1.78e-08
```

We can clearly see that homeowners in zip codes with a high percentage of minorities are taking the default FAIR plan insurance at a higher rate than other zip codes. That is not in doubt. However, can the insurance companies claim that the discrepancy is due to greater risks in some zip codes? The insurance companies could claim that they were denying insurance in neighborhoods where they had sustained large fire-related losses and any discriminatory effect was a by-product of legitimate business practice. We plot some of the variables involved by this question in Figure 11.3:

```
> plot(involact ~ race, chredlin)
> abline(lm(involact ~ race, chredlin))
> plot(fire ~ race, chredlin)
> abline(lm(fire ~ race, chredlin))
```

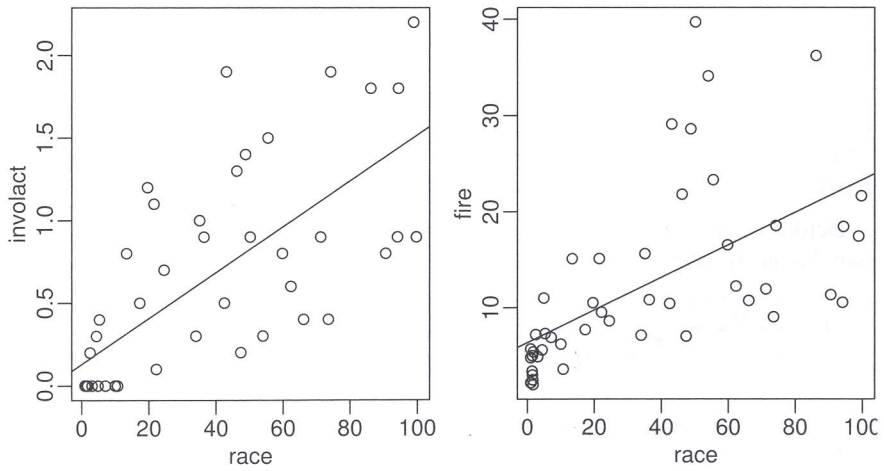

Figure 11.3 *Relationship between `fire`, `race` and `involact` in the Chicago data.*

The question of which variables should also be included in the regression so that their effect may be adjusted for is difficult. Statistically, we can do it, but the important question is whether it should be done at all. For example, it is known that the incomes of women in the United States and other countries are generally lower than those of men. However, if one adjusts for various factors such as type of job and length of service, this gender difference is reduced or can even disappear. The controversy is not statistical but political — should these factors be used to make the adjustment?

For the present data, suppose that the effect of adjusting for income differences was to remove the race effect. This would pose an interesting, but nonstatistical question. I have chosen to include the `income` variable in the analysis just to see what happens.

I have decided to use `log(income)` partly because of skewness in this variable, but also because income is better considered on a multiplicative rather than additive scale. In other words, $1,000 is worth a lot more to a poor person than a millionaire because $1,000 is a much greater fraction of the poor person's wealth.

11.3 Initial Model and Diagnostics

We start with the full model:

```
> g <- lm(involact ~ race + fire + theft + age + log(income),
  chredlin)
> summary(g)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.18554    1.10025   -1.08  0.28755
race         0.00950    0.00249    3.82  0.00045
fire         0.03986    0.00877    4.55  4.8e-05
theft       -0.01029    0.00282   -3.65  0.00073
age          0.00834    0.00274    3.04  0.00413
log(income)  0.34576    0.40012    0.86  0.39254

Residual standard error: 0.335 on 41 degrees of freedom
Multiple R-Squared: 0.752,      Adjusted R-squared: 0.721
F-statistic: 24.8 on 5 and 41 DF,  p-value: 2.01e-11
```

Before leaping to any conclusions, we should check the model assumptions. These two diagnostic plots are seen in Figure 11.4:

```
> plot(fitted(g),residuals(g),xlab="Fitted",ylab="Residuals")
> abline(h=0)
> qqnorm(residuals(g))
> qqline(residuals(g))
```

The diagonal streak in the residual-fitted plot is caused by the large number of zero response values in the data. When $y = 0$, the residual $\hat{\varepsilon} = -\hat{y} = -x^T\hat{\beta}$, hence the line. Turning a blind eye to this feature, we see no particular problem. The Q–Q plot looks fine too.

Now let's look at influence — what happens if points are excluded? We plot the leave-out-one differences in $\hat{\beta}$ for `theft` and the Cook distances:

```
> gi <- influence(g)
> qqnorml(gi$coef[,4])
> halfnorm(cooks.distance(g))
```

See Figure 11.5 where cases 6 and 24 stick out. It is worth looking at other leave-out-one coefficient plots also. We check the jackknife residuals for outliers:

```
> range(rstudent(g))
[1] -3.1850  2.7929
```


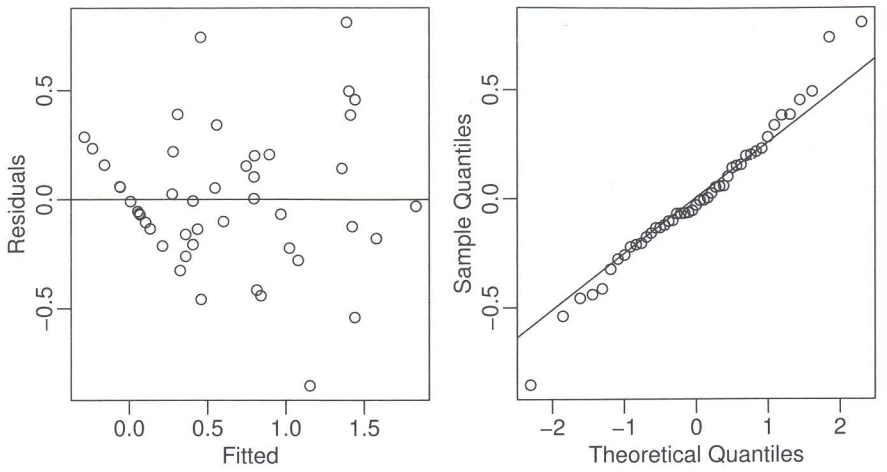

Figure 11.4 *Diagnostic plots of the initial model for the Chicago Insurance data.*

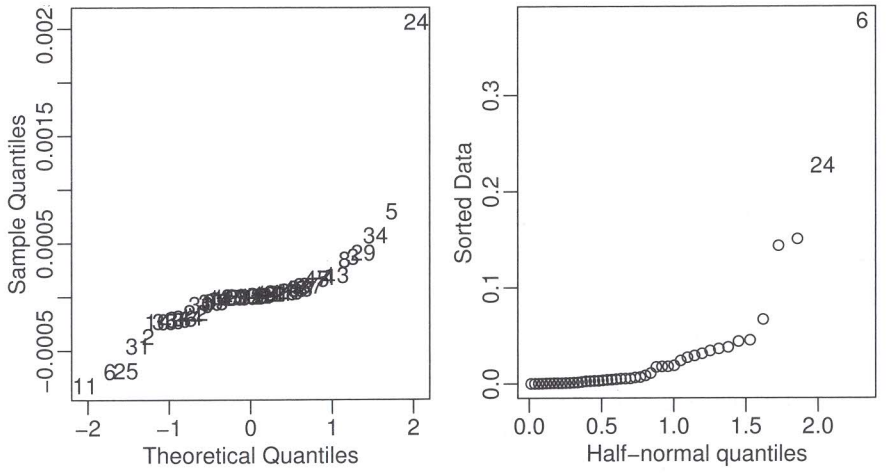

Figure 11.5 *A Q–Q plot of the leave-out-one coefficient differences for the* `theft` *variable is shown on the left. A half-normal plot of the Cook distances is shown on the right.*

There is nothing extreme enough to call an outlier. Let's take a look at the two cases:

```
> chredlin[c(6,24),]
      race fire theft  age involact income side
60610 54.0 34.1    68 52.6      0.3  8.231    n
60607 50.2 39.7   147 83.0      0.9  7.459    n
```

These are high theft and fire zip codes. See what happens when we exclude these points:

```
> g <- lm(involact ~ race + fire + theft + age + log(income),
  chredlin,subset=-c(6,24))
> summary(g)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.57674    1.08005   -0.53    0.596
race         0.00705    0.00270    2.62    0.013
fire         0.04965    0.00857    5.79    1e-06
theft       -0.00643    0.00435   -1.48    0.147
age          0.00517    0.00289    1.79    0.082
log(income)  0.11570    0.40111    0.29    0.775

Residual standard error: 0.303 on 39 degrees of freedom
Multiple R-Squared: 0.804,      Adjusted R-squared: 0.779
F-statistic:   32 on 5 and 39 DF,  p-value: 8.2e-13
```

`theft` and `age` are no longer significant at the 5% level.

11.4 Transformation and Variable Selection

We now look for transformations. We try some partial residual plots as seen in Figure 11.6:

```
> prplot(g,1)
> prplot(g,2)
```

These plots indicate no need to transform. It would have been inconvenient to transform the `race` variable since that would have made interpretation more difficult. Fortunately, we do not need to worry about this. We examined the other partial residual plots and experimented with polynomials for the predictors. No transformation of the predictors appears to be worthwhile.

We choose to avoid even considering a transformation of the response. The zeros in the response would have restricted the possibilities and furthermore would have made interpretation more difficult.

We now move on to variable selection. We are not so much interested in picking one model here because we are mostly interested in the dependency of `involact` on the race variable. So $\hat{\beta}_1$ is the estimate we want to focus on. The problem is that collinearity with the other variables may cause $\hat{\beta}_1$ to vary substantially depending on what other variables are in the model. We address this question here. We leave out the two influential points and force `race` to be included in every model. We do this because race is the primary predictor of interest in this model and we want to

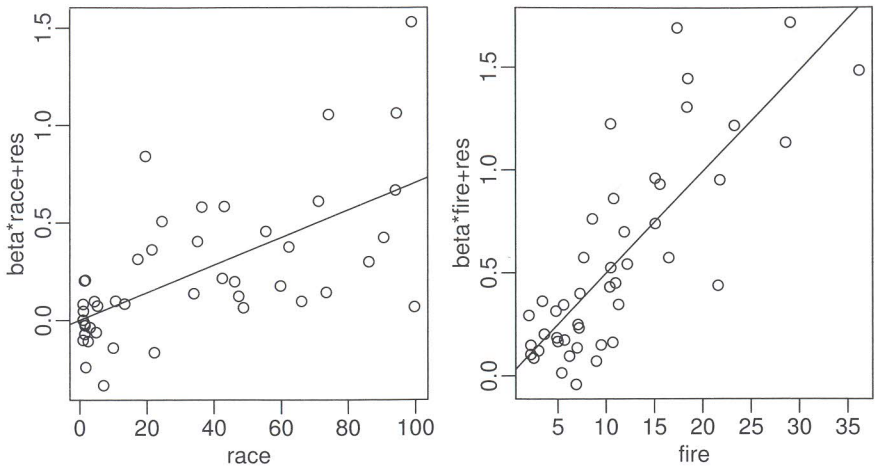

Figure 11.6 *Partial residual plots for* `race` *and* `fire`.

measure its effect. We are not prejudging whether it is significant or not — we will check that later:

```
> chreduc <- chredlin[-c(6,24),]
> library(leaps)
> b<-regsubsets(involact~race + fire + theft + age + log(income),
  force.in=1,data=chreduc)
> (rs <- summary(b))
Subset selection object
            Forced in Forced out
race             TRUE      FALSE
fire            FALSE      FALSE
theft           FALSE      FALSE
age             FALSE      FALSE
log(income)     FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive
         race fire theft age log(income)
2  ( 1 ) "*"  "*"  " "   " " " "
3  ( 1 ) "*"  "*"  " "   "*" " "
4  ( 1 ) "*"  "*"  "*"   "*" " "
5  ( 1 ) "*"  "*"  "*"   "*" "*"
> rs$adj
[1] 0.76855 0.77650 0.78402 0.77895
```

The best model seems to be this one:

```
> g <- lm(involact ~ race + fire + theft + age, chredlin,
  subset=-c(6,24))
> summary(g)
Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.26787    0.13967   -1.92   0.0623
race         0.00649    0.00184    3.53   0.0011
fire         0.04906    0.00823    5.96  5.3e-07
theft       -0.00581    0.00373   -1.56   0.1271
age          0.00469    0.00233    2.01   0.0514

Residual standard error: 0.3 on 40 degrees of freedom
Multiple R-Squared: 0.804,      Adjusted R-squared: 0.784
F-statistic: 40.9 on 4 and 40 DF,  p-value: 1.24e-13
```

The fire rate is significant and actually has higher t-statistics; but nevertheless, we have verified that there is a positive relationship between `involact` and `race` while controlling for a selection of the other variables. Even so we must consider the reliability of this conclusion. For example, would other analysts have come to the same conclusion? One alternative model is:

```
> galt <- lm(involact ~ race+fire+log(income),chredlin,
  subset=-c(6,24))
> summary(galt)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.75326    0.83588    0.90    0.373
race         0.00421    0.00228    1.85    0.072
fire         0.05102    0.00845    6.04  3.8e-07
log(income) -0.36238    0.31916   -1.14    0.263

Residual standard error: 0.309 on 41 degrees of freedom
Multiple R-Squared: 0.786,      Adjusted R-squared: 0.77
F-statistic: 50.1 on 3 and 41 DF,  p-value: 8.87e-14
```

In this model, we see that `race` is not statistically significant. The previous model did fit slightly better, but it is important that there exists a reasonable model in which `race` is not significant since, although the evidence seems fairly strong in favor of a `race` effect, it is not entirely conclusive. Interestingly enough, if `log(income)` is now dropped:

```
> galt <- lm(involact ~ race+fire,chredlin,subset=-c(6,24))
> summary(galt)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.19132    0.08152   -2.35   0.0237
race         0.00571    0.00186    3.08   0.0037
fire         0.05466    0.00784    6.97  1.6e-08

Residual standard error: 0.31 on 42 degrees of freedom
Multiple R-Squared: 0.779,      Adjusted R-squared: 0.769
F-statistic: 74.1 on 2 and 42 DF,  p-value: 1.70e-14
```

we find `race` again becomes significant, which raises again the question of whether income should be adjusted for since it makes all the difference here.

We now return to the two left-out cases. Observe the difference in the fit when the two are reincluded on the best model. The quantities may change but the qualitative message is the same. It is better to include all points if possible, especially in a legal case like this, where excluding points might lead to criticism and suspicion of the results:

```
> g <- lm(involact ~ race + fire + theft + age, chredlin)
> summary(g)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.24312    0.14505   -1.68  0.10116
race         0.00810    0.00189    4.30  0.00010
fire         0.03665    0.00792    4.63  3.5e-05
theft       -0.00959    0.00269   -3.57  0.00092
age          0.00721    0.00241    2.99  0.00460

Residual standard error: 0.334 on 42 degrees of freedom
Multiple R-Squared: 0.747,      Adjusted R-squared: 0.723
F-statistic:   31 on 4 and 42 DF,  p-value: 4.8e-12
```

The main message of the data is not changed. On checking the diagnostics, I found no trouble. So it looks like there is moderately good evidence that zip codes with high minority populations are being "redlined." While there is evidence that some of the relationship between `race` and `involact` can be explained by the fire rate, there is still a component that cannot be attributed to the other variables.

11.5 Discussion

There is some ambiguity in the conclusion here. These reservations have several sources.

There is some doubt because the response is not a perfect measure of people being denied insurance. It is an aggregate measure that raises the problem of ecological correlations. We have implicitly assumed that the probability a minority homeowner would obtain a FAIR plan after adjusting for the effect of the other covariates is constant across zip codes. This is unlikely to be true. If the truth is simply a variation about some constant, then our conclusions will still be reasonable, but if this probability varies in a systematic way, then our conclusions may be off the mark. It would be a very good idea to obtain some individual level data.

Another point to be considered is the size of the effect. The largest value of the response is only 2.2% and most other values are much smaller. Even assuming the worst, the number of people affected is small.

There is also the problem of a potential latent variable that might be the true cause of the observed relationship. Someone with firsthand knowledge of the insurance business might propose one. This possibility always casts a shadow of doubt on our conclusions.

Another issue that arises in cases of this nature is how much the data should be aggregated. For example, suppose we fit separate models to the two halves of the city. Fit the model to the south of Chicago:

```
> g <- lm(involact ~ race+fire+theft+age, subset=(side == "s"),
  chredlin)
> summary(g)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.23441    0.23774   -0.99    0.338
race         0.00595    0.00328    1.81    0.087
fire         0.04839    0.01689    2.87    0.011
theft       -0.00664    0.00844   -0.79    0.442
age          0.00501    0.00505    0.99    0.335

Residual standard error: 0.351 on 17 degrees of freedom
Multiple R-Squared: 0.743,      Adjusted R-squared: 0.683
F-statistic: 12.3 on 4 and 17 DF,  p-value: 6.97e-05
```

and now to the north:

```
> g <- lm(involact ~ race+fire+theft+age, subset=(side == "n"),
  chredlin)
> summary(g)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.31857    0.22702   -1.40    0.176
race         0.01256    0.00448    2.81    0.011
fire         0.02313    0.01398    1.65    0.114
theft       -0.00758    0.00366   -2.07    0.052
age          0.00820    0.00346    2.37    0.028

Residual standard error: 0.343 on 20 degrees of freedom
Multiple R-Squared: 0.756,      Adjusted R-squared: 0.707
F-statistic: 15.5 on 4 and 20 DF,  p-value: 6.52e-06
```

We see that `race` is significant in the north, but not in the south. By dividing the data into smaller and smaller subsets it is possible to dilute the significance of any predictor. On the other hand, it is important not to aggregate all data without regard to whether it is reasonable. Clearly a judgment has to be made and this can be a point of contention in legal cases.

There are some special difficulties in presenting this during a court case. With scientific inquiries, there is always room for uncertainty and subtlety in presenting the results, particularly if the subject matter is not contentious. In an adversarial proceeding, it is difficult to present statistical evidence when the outcome is not clear-cut, as in this example. There are particular difficulties in explaining such evidence to nonmathematically trained people.

After all this analysis, the reader may be feeling somewhat dissatisfied. It seems we are unable to come to any truly definite conclusions and everything we say has been hedged with "ifs" and "buts." Winston Churchill once said:

> Indeed, it has been said that democracy is the worst form of Government except all those other forms that have been tried from time to time.

We might say the same about statistics with respect to how it helps us reason in the face of uncertainty. It is not entirely satisfying but the alternatives are worse.

CHAPTER 12

Missing Data

Missing data occur when some values of some cases are missing. This is not uncommon. Dealing with missing data is time consuming. Fixing up problems caused by missing data sometimes takes longer than the analysis.

What can be done? Obviously, finding the missing values is the best option, but this is not always possible. Next, ask why the data are missing. If the reason an observation is missing is noninformative, then a fix is easier. For example, if a data point is missed because it was large in value, then this could cause some bias and a simple fix is not possible. Patients may drop out of a drug study, because they believe their treatment is not working — this would cause bias.

Here are several fix-up methods to use when data are missing for noninformative reasons:

1. Delete the case with missing observations. This is OK if this only causes the loss of a relatively small number of cases. This is the simplest solution.
2. Fill in or *impute* the missing values. Use the rest of the data to predict the missing values. Simply replacing the missing value of a predictor with the average value of that predictor is one easy method. Using regression on the other predictors is another possibility. It is not clear how much the diagnostics and inference on the filled-in dataset are affected. Some additional uncertainty is caused by the imputation, which needs to be taken into account. Multiple imputation can capture some of this uncertainty.
3. Consider just (x_i, y_i) pairs with some observations missing. The means and SDs of x and y can be used in the estimate even when a member of a pair is missing. An analogous method is available for regression problems. This is called the *missing value correlation* method.
4. Maximum likelihood methods can be used assuming the multivariate normality of the data. The EM algorithm is often used here. We will not explain the details, but the idea is essentially to treat missing values as nuisance parameters.

Suppose some of the values in the Chicago Insurance dataset were missing. I randomly declared some of the observations missing in this modified dataset. Read it in and take a look:

```
> data(chmiss)
> chmiss
      race fire theft  age involact income
60626 10.0  6.2    29 60.4       NA 11.744
60640 22.2  9.5    44 76.5      0.1  9.323
60613 19.6 10.5    36   NA      1.2  9.948
```